DNASTAR®
Software for life scientists

SeqMan NGen 12.2 vs. Two BWA+GATK Workflows: Variant Detection Comparison Using Illumina Data from NA12878

To obtain reliable variant results, the accuracy of sequence alignment, consensus calling and variant detection is of paramount importance. Throughout its history, DNASTAR has emphasized the development of exceptionally accurate software, ensuring that users will obtain the highest quality results. To assess the accuracy of DNASTAR's next-generation sequence aligner and variant caller, we compared whole exome results from DNASTAR's SeqMan NGen 12.2 with those from two other industry-leading pipelines:

- The Burrows-Wheeler Aligner (BWA) read-mapper in combination with the Broad Institute's Genome Analysis Toolkit (GATK) Unified Genotyper variant caller
- The BWA read-mapper in combination with the GATK Haplotype Base Caller

Our results demonstrate that SeqMan NGen 12.2 achieves better sensitivity than either of the BWA+GATK workflows. SeqMan NGen also aligns the data and performs variant calling an average of 5 times faster than the BWA+GATK pipelines.

Input Data

All three software pipelines used a common set of input data derived from the HapMap/1000 CEU female, NA12878. Through the Genome in a Bottle Consortium (GIAB), the National Institute of Standards and Technology (NIST) has developed a highly accurate and well-characterized set of genome-wide reference materials[1] for NA12878, including BED and VCF files of high-quality sequence regions and variant calls, respectively. The GIAB call sets were built from the integration of eleven NA12878 whole human genome data sets and three exome data sets, generated across five sequencing platforms to eliminate bias from any single platform. These data can be used as a benchmark when assessing variant call accuracy.

Input data used for this study are described below:

- The human genome reference sequence GRCh37 (hg19).
- Illumina paired-end read data. Evaluations were made for several whole exome data sets derived from NA12878 (Table 1).

  Table 1. Data Sets

Nickname	Data Source	Download Link
Garvan	Illumina exome data from the Garvan Institute of Medical Research	Download
ARUP	Illumina exome data from NCBI's Short Read Archive data set "ARUP exome sequencing of human HapMap individual NA12878," submitted by ARUP Laboratories (Salt Lake City, UT).	Download
Icahn	NCBI's Short Read Archive Illumina exome data set "High-coverage whole exome sequencing of CEPH/UTAH female individual (HapMap: NA12878)," submitted by the Icahn School of Medicine at Mount Sinai (New York, NY).	Download

- A BED file for each data set, detailing the targeted regions of interest. Each BED file was constructed by intersecting the GIAB (version 2.19) high confidence region BED file for NA12878 with the appropriate exome target region BED file derived from the capture manifest file. Resulting continuous sequence fragments less than 20 bases in length were not used.

- For each data set, the VCF file containing the GIAB high confidence variant calls for the NA12878 genome was intersected with the corresponding BED file created above. This removes variants that are not contained in the intervals for the data set's capture region, and which would otherwise be scored as "negative" calls.

We thank the researchers for their generosity in making these data publicly available.

Software Workflows

All alignment and variant detection was performed on the same 64-bit Mac Pro 6 workstation (with OS X 10.10) using each application's default settings for high-sensitivity alignment. In order to simplify timing comparisons, only one job was run at a time, with no other applications running. For each experiment, an individual data set was aligned against the entire human genome reference sequence, GRCh37 (hg19).

- DNASTAR's SeqMan NGen algorithm analyzes fully gapped alignments in-stream using a modified version of the MAQ variant caller[2] to produce variant and reference call files for each position in the intersected BED file for that experiment.
- The BWA+GATK workflows used BWA[3] for the alignment stage. Alignment results were then processed with the GATK RealignerTargetCreator and GATK IndelRealigner, followed by variant detection with either the GATK Unified Genotyper or the GATK Haplotype Base Caller. The workflow utilizing BWA and the Unified Genotyper is based on the Illumina MiSeq Reporter variant detection pipeline. These workflows use the SAMtools[4] utility program for various procedures such as converting alignments from SAM to BAM format, removing duplicates, and sorting and indexing BAM files. The equivalent functionality is built directly into SeqMan NGen. Default settings were used for all components other than the GATK variant callers. For the GATK variant callers, a call quality threshold of Q=10 was used to match the default PnotRef setting for SeqMan NGen.

All workflows utilized default filters to focus the variant analysis on positions with reasonable data support, based on 1) the probability that the called base is not the homozygous reference base and 2) a specified minimum depth of coverage (Table 2). All positions meeting the coverage requirement but not the probability requirement were classified as reference calls.

Table 2. Default Filters Used in Each Workflow

Software	Default Settings		
	Minimum variant frequency	Probability variant is not homozygous reference base	Minimum depth of coverage
DNASTAR's SeqMan NGen 12.2	15%	Minimum PNotRef = 90%	20
BWA+GATK (both workflows)	N/A	Q=10 (equivalent to PNotRef = 90%)	20

Calculations

After aligning the data using DNASTAR's SeqMan NGen or one of the BWA+GATK pipelines, Perl scripts were used to independently compare the accuracy of variant detection results (substitutions, and insertions and deletions of up to 10 bp) relative to the "answer" provided by GIAB. To accomplish this task, the VCF files from GIAB were compared to either the variant reports exported from SeqMan Pro® or the VCF files produced by GATK. Each position was then placed into one of four categories (Table 3, unshaded rows) and used to calculate a series of three statistical metrics (shaded rows).

Table 3. Accuracy Metrics

Column Name	Description
Sensitivity	The proportion of true positives that are correctly identified: TP / (TP + FN)
Specificity	The proportion of true negatives that are correctly identified: TN / (TN + FP)
False Discovery Rate (FDR)	The proportion of false positives among all discoveries: FP / (TP + FP)
True Positives (TP)	Called variants with a corresponding position in the GIAB VCF file
False Positives (FP)	Called variants without a corresponding position in the VCF file
True Negatives (TN)	Called reference bases without a corresponding position in the VCF file
False Negatives (FN)	Called reference bases with a corresponding position in the VCF file

Results

The accuracy of variant detection relative to the "answer" provided by GIAB was calculated for each workflow (Table 4).

Table 4. Accuracy Results for NA12878 Exome Data

Data Set	Workflow (Mapper/Variant Caller)	Sensitivity	Specificity	FDR	TP	FP	TN	FN	Elapsed Time*
Garvan	DNASTAR's SeqMan NGen 12.2	99.56%	99.999%	1.29%	15,272	200	24,882,436	67	1.3 hr
	BWA/GATK Unified Genotyper	99.09%	99.999%	1.08%	15,161	166	24,798,616	139	6.0 hr
	BWA/GATK Haplotype Base Caller	99.14%	99.999%	0.97%	15,168	149	24,798,633	132	6.3 hr
ARUP	DNASTAR's SeqMan NGen 12.2	99.51%	99.9997%	0.53%	15,009	80	24,532,535	74	0.5 hr
	BWA/GATK Unified Genotyper	99.00%	99.9998%	0.32%	14,957	48	24,542,654	151	2.7 hr
	BWA/GATK Haplotype Base Caller	99.14%	99.9998%	0.39%	14,978	59	24,542,643	130	2.8 hr
Icahn	DNASTAR's SeqMan NGen 12.2	99.49%	99.999%	1.14%	36,464	422	43,817,958	186	1.1 hr
	BWA/GATK Unified Genotyper	98.81%	99.9988%	1.38%	35,988	503	43,546,321	432	4.9 hr
	BWA/GATK Haplotype Base Caller	99.01%	99.9987%	1.60%	36,061	586	43,546,238	359	5.0 hr

* Includes times to index genomes or build SeqMan NGen template mers.

Conclusion

The results presented in Table 4 demonstrate that SeqMan NGen 12.2 is a fast, high accuracy read-mapper/variant caller for Illumina whole exome sequencing data. SeqMan NGen provides higher sensitivity for each data set tested, compared to the BWA+GATK pipelines currently in use. In addition, SeqMan NGen analyzes variants in whole exome data an average of 5 times faster than the BWA+GATK workflows.

Other benefits to using SeqMan NGen include its Graphical User Interface (GUI), its availability on multiple platforms (Linux, Windows and Macintosh), and the ease of installing and using the software. With SeqMan NGen, jobs can be running within minutes of the initial download. In addition, SeqMan NGen works in tandem with SeqMan Pro and ArrayStar for fully integrated analysis as part of the Lasergene Genomics Suite.

To replicate these calculations, please download a free trial of Lasergene 12 and utilize the cited, publicly accessible input data.

References

1) Zook J, *et al* (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology 32, 246-251.
2) Li H, *et al* (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851-1858.
3) Li H and Durbin R (2009). Fast and Accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 15, 1754-1760.
4) Li H, *et al.* (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.